WHAT IS CLAIMED IS:

1. A method of extracting significant patterns from a dataset having a plurality of sequences defined over a lexicon of tokens, the method comprising, for each sequence of the plurality of sequences: searching for partial overlaps between said sequence and other sequences of the dataset, applying a significance test on said partial overlaps, and defining a most significant partial overlap as a significant pattern of said sequence, thereby extracting significant patterns from the dataset.

2. The method of claim 1, wherein said search for partial overlaps is by constructing a graph having a plurality of paths representing the dataset and searching for partial overlaps between paths of said graph.

3. The method of claim 2, wherein said search for partial overlaps between paths of said graph comprises:

defining, for each path, a set of sub-paths of variable lengths, thereby defining a plurality of sets of sub-paths; and

for each set of sub-paths, comparing each sub-path of said set with sub-paths of other sets.

4. The method of claim 2, wherein said graph comprises a plurality of vertices, each representing one token of the lexicon, and further wherein each path of said plurality of paths comprises a sequence of vertices respectively corresponding to one sequence of the dataset.

5. The method of claim 2, further comprising calculating, for each path, a set of probability functions characterizing said partial overlaps.

6. The method of claim 5, wherein said application of said significance test is by evaluating a statistical significance of said set of probability functions.

7. The method of claim 1, further comprising grouping at least a few tokens of said significant pattern, thereby redefining the dataset.

8. The method of claim 1, wherein the dataset comprises a corpus of text.

9. The method of claim 1, wherein the dataset comprises a protein database.

10. The method of claim 1, wherein the dataset comprises a protein database.

11. The method of claim 1, wherein the dataset comprises a DNA database.

12. The method of claim 1, wherein the dataset comprises an RNA database.

13. The method of claim 1, wherein the dataset comprises a recorded speech.

14. The method of claim 1, wherein the dataset comprises a corpus of music notes.

15. The method of claim 1, wherein the dataset comprises a weblog database.

16. The method of claim 1, wherein the dataset comprises trajectory records of a transportation network.

17. The method of claim 1, wherein the dataset comprises activity records of a self-active system.

18. The method of claim 1, wherein the dataset comprises records of operational steps in a technical process.

19. A method of generalizing a dataset having a plurality of sequences defined over a lexicon of tokens, the method comprising:

searching over the dataset for similarity sets, each similarity set comprising a plurality of segments of size L having L-S common tokens and S uncommon tokens, each of said plurality of segments being a portion of a different sequence of the dataset; and

defining a plurality of equivalence classes corresponding to uncommon tokens of at least one similarity set, thereby generalizing the dataset.

20. The method of claim 19, wherein said definition of said plurality of equivalence classes comprises, for each segment of each similarity set:

extracting a significant pattern corresponding to a most significant partial overlap between said segment and other segments or combination of segments of said similarity set, thereby providing, for each similarity set, a plurality of significant patterns; and

using said plurality of significant patterns for classifying tokens of said similarity set into at least one equivalence class;

thereby defining said plurality of equivalence classes.

21. The method of claim 20, wherein said classification of said tokens comprises, selecting a leading significant pattern of said similarity set, and defining uncommon tokens of segments corresponding to said leading significant pattern as an equivalence class.

22. The method of claim 20, further comprising, prior to said search for said similarity sets:

extracting a plurality of significant patterns from the dataset, each significant pattern of said plurality of significant patterns corresponding to a most significant partial overlap between one sequence of the dataset and other sequences of the dataset; and

for each significant pattern of said plurality of significant patterns, grouping at least a few tokens of said significant pattern, thereby redefining the dataset.

23. The method of claim 19, further comprising, for each similarity set having at least one equivalence class, grouping at least a few tokens of said similarity set thereby redefining the dataset.

24. The method of claim 22, further comprising, for each similarity set having at least one equivalence class, grouping at least a few tokens of said similarity set thereby redefining the dataset.

25. The method of claim 19, further comprising for each sequence, searching over said sequence for tokens being identified as members of previously defined equivalence classes, and attributing a respective equivalence class to each identified token, thereby generalizing said sequence, thereby further generalizing the dataset.

26. The method of claim 25, wherein said attribution of said respective equivalence class to said identified token is subjected to a generalization test.

27. The method of claim 26, wherein said generalization test comprises determining a number of different sequences having tokens being identified as other elements of said respective equivalence class, and if said number of different sequences is larger than a predetermined generalization threshold, then attributing said respective equivalence class to said identified token.

28. The method of claim 25, wherein said attribution of said respective equivalence class to said identified token is subjected to a significance test.

29. The method of claim 28, wherein said significance test comprises:

for each sequence having elements of said respective equivalence class, searching for partial overlaps between said sequence and other sequences having elements of said respective equivalence class, and defining a most significant partial overlap as a significant pattern of said sequence, thereby extracting a plurality of significant patterns;

selecting a leading significant pattern of said plurality of significant patterns; and

if said leading significant pattern includes said identified token, then attributing said respective equivalence class to said identified token.

30. The method of claim 22, further comprising constructing a graph having a plurality of paths representing the dataset, wherein each extraction of significant pattern is by searching for partial overlaps between paths of said graph.

31. The method of claim 30, wherein said graph comprises a plurality of vertices, each representing one token of the lexicon, and further wherein each path of said plurality of paths comprises a sequence of vertices respectively corresponding to one sequence of the dataset.

32. The method of claim 30, further comprising calculating, for each path, a set of probability functions characterizing said partial overlaps.

33. The method of claim 32, wherein said most significant partial overlap is determined by a significance test being performed by evaluating a statistical significance of said set of probability functions.

34. The method of claim 19, wherein the dataset comprises a corpus of text.

35. The method of claim 19, wherein the dataset comprises a protein database.

36. The method of claim 19, wherein the dataset comprises a protein database.

37. The method of claim 19, wherein the dataset comprises a DNA database.

38. The method of claim 19, wherein the dataset comprises an RNA database.

39. The method of claim 19, wherein the dataset comprises a recorded speech.

40. The method of claim 19, wherein the dataset comprises a corpus of music notes.

41. The method of claim 19, wherein the dataset comprises a weblog database.

42. The method of claim 19, wherein the dataset comprises trajectory records of a transportation network.

43. The method of claim 19, wherein the dataset comprises activity records of a self-active system.

44. The method of claim 19, wherein the dataset comprises records of operational steps in a technical process.

45. A method of extracting significant patterns from a dataset having a plurality of sequences defined over a lexicon of tokens, the method comprising:

(a) constructing a graph having a plurality of vertices and paths of vertices, each vertex representing one token of the lexicon, such that each sequence of the plurality of sequences is represented by one path of said plurality of paths; and

(b) for each path of said plurality of paths: searching for partial overlaps between said path and other paths, applying a significance test on said partial overlaps, and defining a most significant partial overlap as a significant pattern of said path;

thereby extracting significant patterns from the dataset.

46. The method of claim 45, wherein said search for partial overlaps between paths of said graph comprises defining a set of sub-paths of variable lengths for said path, and comparing each sub-path of said path with sub-paths of other paths.

47. The method of claim 45, further comprising marking endpoints of each path of said plurality of paths, by adding a first marking vertex before a first vertex of said path and a second marking vertex after a last vertex of said path.

48. The method of claim 46, further comprising calculating, for each path, a set of probability functions characterizing said partial overlaps.

49. The method of claim 48, wherein said application of said significance test is by evaluating a statistical significance of said set of probability functions.

50. The method of claim 49, wherein said set of probability functions constitutes a variable-order Markov matrix.

51. The method of claim 50, wherein said evaluation of said statistical significance is by using elements of said variable-order Markov matrix to calculate a set of cohesion coefficients for each path, and selecting a supremum of said set of cohesion coefficients.

52. The method of claim 49, wherein said set of probability functions comprises for each sub-path of said path, a probability function characterizing a rightward direction on said sub-path, and a probability function characterizing a leftward direction on said sub-path.

53. The method of claim 45, further comprising for each significant pattern, defining a pattern-vertex representing at least a few vertices of said significant pattern, thereby redefining said graph.

54. The method of claim 45, wherein the dataset comprises a corpus of text.

55. The method of claim 45, wherein the dataset comprises a protein database.

56. The method of claim 45, wherein the dataset comprises a protein database.

57. The method of claim 45, wherein the dataset comprises a DNA database.

58. The method of claim 45, wherein the dataset comprises an RNA database.

59. The method of claim 45, wherein the dataset comprises a recorded speech.

60. The method of claim 45, wherein the dataset comprises a corpus of music notes.

61. The method of claim 45, wherein the dataset comprises a weblog database.

62. The method of claim 45, wherein the dataset comprises trajectory records of a transportation network.

63. The method of claim 45, wherein the dataset comprises activity records of a self-active system.

64. The method of claim 45, wherein the dataset comprises records of operational steps in a technical process.

65. A method of generalizing a dataset having a plurality of sequences defined over a lexicon of tokens, the method comprising:

(a) constructing a graph having a plurality of vertices and paths of vertices, each vertex representing one token of the lexicon, such that each sequence of the plurality of sequences is represented by one path of said plurality of paths;

(b) searching over said plurality of paths for similarity sets, each similarity set comprising a plurality of paths sharing L-S vertices within an L-size window, hence defining S slots each being a set of different vertices; and

(c) defining a plurality of equivalence classes corresponding to at least one slot of at least one similarity set;

thereby generalizing the dataset.

66. The method of claim 65, wherein said definition of said plurality of equivalence classes comprises, for each segment of each similarity set:

for each path of each similarity set, extracting a significant pattern corresponding to a most significant partial overlap between said path and other paths or combinations of paths of said similarity set, thereby providing, for each similarity set, a plurality of significant patterns; and

using said plurality of significant patterns for classifying vertices of said similarity set into at least one equivalence class;

thereby defining said plurality of equivalence classes.

67. The method of claim 66, wherein said classification of vertices comprises selecting a leading significant pattern of said similarity set, and defining a slot corresponding to said leading significant pattern as an equivalence class.

68. The method of claim 66, further comprising redefining said graph prior to said step (b), said redefinition of said graph comprising:

for each path of said plurality of paths, extracting a significant pattern corresponding to a partial overlap between said path and paths other then said path, thereby providing a plurality of significant patterns; and

for each significant pattern of said plurality of significant patterns, defining a pattern-vertex representing at least a few vertices of said significant pattern..

69. The method of claim 65, further comprising, subsequently to said step (c), defining, for each similarity set having at least one equivalence class, a generalized-vertex representing all vertices of a respective L-size window of said similarity set, thereby redefining said graph.

70. The method of claim 68, further comprising, subsequently to said step (c), defining, for each similarity set having at least one equivalence class, a generalized-vertex representing all vertices of a respective L-size window of said similarity set, thereby redefining said graph.

71. The method of claim 69, further comprising repeating said step (b) and said step (c), subsequently to said redefinition of said graph, at least once.

72. The method of claim 69, further comprising repeating said step (b) and said step (c), subsequently to said redefinition of said graph, a plurality of times.

73. The method of claim 72, further comprising repeating said steps (b) and said step (c), a plurality of times while permuting a searching order of said step (b), thereby providing a plurality of generalized datasets, each characterized by a generalization factor, and selecting a generalized dataset corresponding to a maximal generalization factor.

74. The method of claim 73, wherein said generalization factor is defined as a ratio between a number of sequences of said generalized dataset and a number of sequences of the dataset.

75. The method of claim 72, further comprising repeating said steps (b) and said step (c), a plurality of times while permuting a searching order of said step (b), thereby providing a plurality of generalized datasets, each characterized by a precision value and a recall value, and selecting a generalized dataset corresponding to an optimal combination of said precision value and said recall value.

76. The method of claim 65, further comprising for each path, searching over said path for vertices being identified as members of previously defined equivalence classes, and attributing a respective equivalence class to each identified vertex, thereby generalizing said path, thereby further generalizing the dataset.

77. The method of claim 76, wherein said attribution of said respective equivalence class to said identified vertex is subjected to a generalization test.

78. The method of claim 77, wherein said generalization test comprises determining a number of different paths having, within said L-size window, vertices being identified as other elements of said respective equivalence class, and if said number of different paths is larger than a predetermined generalization threshold, then attributing said respective equivalence class to said identified vertex.

79. The method of claim 76, wherein said attribution of said respective equivalence class to said identified vertex is subjected to a significance test.

80. The method of claim 79, wherein said significance test comprises:

for each path having elements of said respective equivalence class, searching for partial overlaps between said path and other paths having elements of said respective equivalence class, and defining a most significant partial overlap as a significant pattern of said path, thereby extracting a plurality of significant patterns;

selecting a leading significant pattern of said plurality of significant patterns; and

if said leading significant pattern includes said identified vertex, then attributing said respective equivalence class to said identified vertex.

81. The method of claim 65, further comprising marking endpoints of each path of said plurality of paths, by adding a first marking vertex before a first vertex of said path and a second marking vertex after a last vertex of said path.

82. The method of claim 68, further comprising calculating, for each path, a set of probability functions characterizing said partial overlaps.

83. The method of claim 77, wherein said extraction of said significant pattern from said path is by a evaluating a statistical significance of said set of probability functions.

84. The method of claim 65, wherein the dataset comprises a corpus of text.

85. The method of claim 65, wherein the dataset comprises a protein database.

86. The method of claim 65, wherein the dataset comprises a protein database.

87. The method of claim 65, wherein the dataset comprises a DNA database.

88. The method of claim 65, wherein the dataset comprises an RNA database.

89. The method of claim 65, wherein the dataset comprises a recorded speech.

90. The method of claim 65, wherein the dataset comprises a corpus of music notes.

91. The method of claim 65, wherein the dataset comprises a weblog database.

92. The method of claim 65, wherein the dataset comprises trajectory records of a transportation network.

93. The method of claim 65, wherein the dataset comprises activity records of a self-active system.

94. The method of claim 65, wherein the dataset comprises records of operational steps in a technical process.

95. A method of executing at least one action based on at least one instruction, the method comprising, inputting a dataset having a plurality of sequences defined over a lexicon of tokens, learning said dataset so as to provide a generalized dataset, inputting an instruction, using said generalized dataset for determining an action corresponding to said instruction, and executing said action;

wherein said learning said dataset comprises:

(a) constructing a graph having a plurality of vertices and paths of vertices, each vertex representing one token of the lexicon, such that each sequence of the plurality of sequences is represented by one path of said plurality of paths;

(b) searching over said plurality of paths for similarity sets, each similarity set comprising a plurality of paths sharing L-S vertices within an L-size window, hence defining S slots each being a set of different vertices; and

(c) defining a plurality of equivalence classes corresponding to at least one slot of at least one similarity set, thereby providing a generalized dataset.

96. The method of claim 95, wherein said input of said instruction, said use of said generalized dataset for determining said action, and said execution of said action is repeated at least once.

97. The method of claim 95, wherein said instruction is a written instruction.

98. The method of claim 95, wherein said instruction is a verbal instruction.

99. The method of claim 95, wherein said definition of said plurality of equivalence classes comprises, for each segment of each similarity set:

for each path of each similarity set, extracting a significant pattern corresponding to a most significant partial overlap between said path and other paths

or combinations of paths of said similarity set, thereby providing, for each similarity set, a plurality of significant patterns; and

using said plurality of significant patterns for classifying vertices of said similarity set into at least one equivalence class;

thereby defining said plurality of equivalence classes.

100. The method of claim 99, wherein said classification of vertices comprises selecting a leading significant pattern of said similarity set, and defining a slot corresponding to said leading significant pattern as an equivalence class.

101. The method of claim 99, further comprising redefining said graph prior to said step (b), said redefinition of said graph comprising:

for each path of said plurality of paths, extracting a significant pattern corresponding to a partial overlap between said path and paths other then said path, thereby providing a plurality of significant patterns; and

for each significant pattern of said plurality of significant patterns, defining a pattern-vertex representing at least a few vertices of said significant pattern..

102. The method of claim 95, further comprising, subsequently to said step (c), defining, for each similarity set having at least one equivalence class, a generalized-vertex representing all vertices of a respective L-size window of said similarity set, thereby redefining said graph.

103. The method of claim 101, further comprising, subsequently to said step (c), defining, for each similarity set having at least one equivalence class, a generalized-vertex representing all vertices of a respective L-size window of said similarity set, thereby redefining said graph.

104. The method of claim 102, further comprising repeating said step (b) and said step (c), subsequently to said redefinition of said graph, at least once.

105. The method of claim 102, further comprising repeating said step (b) and said step (c), subsequently to said redefinition of said graph, a plurality of times.

106. The method of claim 95, further comprising for each path, searching over said path for vertices being identified as members of previously defined equivalence classes, and attributing a respective equivalence class to each identified vertex, thereby generalizing said path, thereby further generalizing the dataset.

107. The method of claim 106, wherein said attribution of said respective equivalence class to said identified vertex is subjected to a generalization test.

108. The method of claim 107, wherein said generalization test comprises determining a number of different paths having, within said L-size window, vertices being identified as other elements of said respective equivalence class, and if said number of different paths is larger than a predetermined generalization threshold, then attributing said respective equivalence class to said identified vertex.

109. The method of claim 106, wherein said attribution of said respective equivalence class to said identified vertex is subjected to a significance test.

110. The method of claim 109, wherein said significance test comprises:

for each path having elements of said respective equivalence class, searching for partial overlaps between said path and other paths having elements of said respective equivalence class, and defining a most significant partial overlap as a significant pattern of said path, thereby extracting a plurality of significant patterns;

selecting a leading significant pattern of said plurality of significant patterns; and

if said leading significant pattern includes said identified vertex, then attributing said respective equivalence class to said identified vertex.

111. The method of claim 95, further comprising marking endpoints of each path of said plurality of paths, by adding a first marking vertex before a first vertex of said path and a second marking vertex after a last vertex of said path.

112. The method of claim 101, further comprising calculating, for each path, a set of probability functions characterizing said partial overlaps.

113. The method of claim 107, wherein said extraction of said significant pattern from said path is by a evaluating a statistical significance of said set of probability functions.

114. The method of claim 95, wherein the dataset comprises a corpus of text.

115. The method of claim 95, wherein the dataset comprises a recorded speech.

116. An apparatus for extracting significant patterns from a dataset having a plurality of sequences defined over a lexicon of tokens, the apparatus comprising:

(a) a searcher, for searching for partial overlaps between said sequence and other sequences of the dataset;

(b) a testing unit, for applying a significance test on said partial overlaps; and

(c) a definition unit, for defining a most significant partial overlap as a significant pattern of said sequence.

117. The apparatus of claim 116, further comprising a constructor, for constructing a graph having a plurality of paths representing the dataset.

118. The apparatus of claim 117, wherein said searcher is designed to search for partial overlaps between paths of said graph.

119. The apparatus of claim 117, wherein said searcher comprises:

a sub-path definer, for defining a plurality of sets of sub-paths, one sets of sub-path for each path; and

a sub-path comparer, for comparing for a given set of sub-paths, each sub-path of said set with sub-paths of other sets.

120. The apparatus of claim 118, wherein said graph comprises a plurality of vertices, each representing one token of the lexicon, and further wherein each path of

said plurality of paths comprises a sequence of vertices respectively corresponding to one sequence of the dataset.

121. The apparatus of claim 118, further comprising electronic-calculation functionality for calculating, for each path, a set of probability functions characterizing said partial overlaps.

122. The apparatus of claim 121, wherein said testing unit is capable of evaluating a statistical significance of said set of probability functions.

123. The apparatus of claim 116, wherein the dataset comprises a corpus of text.

124. The apparatus of claim 116, wherein the dataset comprises a protein database.

125. The apparatus of claim 116, wherein the dataset comprises a protein database.

126. The apparatus of claim 116, wherein the dataset comprises a DNA database.

127. The apparatus of claim 116, wherein the dataset comprises an RNA database.

128. The apparatus of claim 116, wherein the dataset comprises a recorded speech.

129. The apparatus of claim 116, wherein the dataset comprises a corpus of music notes.

130. The method of claim 116, wherein the dataset comprises a weblog database.

131. The method of claim 116, wherein the dataset comprises trajectory records of a transportation network.

132. The method of claim 116, wherein the dataset comprises activity records of a self-active system.

133. The method of claim 116, wherein the dataset comprises records of operational steps in a technical process.

134. An apparatus for generalizing a dataset having a plurality of sequences defined over a lexicon of tokens, the apparatus comprising:

(a) a searcher, for searching over the dataset for similarity sets, each similarity set comprising a plurality of segments of size L having L-S common tokens and S uncommon tokens, each of said plurality of segments being a portion of a different sequence of the dataset; and

(b) a definition unit, for defining a plurality of equivalence classes corresponding to uncommon tokens of at least one similarity set, thereby generalizing the dataset.

135. The apparatus of claim 134, further comprising an extractor, capable of extracting, for a given set of sequences, a significant pattern corresponding to a most significant partial overlap between one sequence of said set of sequences and other sequences of said set of sequences, thereby providing, for said given set of sequences, a plurality of significant patterns.

136. The apparatus of claim 135, wherein said given set of sequences is a similarity set, hence said plurality of significant patterns corresponds to said similarity set.

137. The apparatus of claim 136, wherein said definition unit comprises a classifier, capable of classifying tokens of said similarity set into at least one equivalence class using said plurality of significant patterns.

138. The apparatus of claim 135, wherein said classifier is designed for selecting a leading significant pattern of said similarity set, and defining uncommon tokens of segments corresponding to said leading significant pattern as an equivalence class.

139. The apparatus of claim 135, wherein said given set of sequences is the dataset, hence said plurality of significant patterns corresponds to the dataset.

140. The apparatus of claim 135, further comprising a first grouper for grouping at least a few tokens of each significant pattern of said plurality of significant patterns.

141. The apparatus of claim 140, further comprising a second grouper, for grouping at least a few tokens of each similarity set having at least one equivalence class.

142. The apparatus of claim 134, further comprising a second definition unit having a second searcher, for searching over each sequence for tokens being identified as members of previously defined equivalence classes, wherein said second definition unit is designed to attribute a respective equivalence class to each identified token.

143. The apparatus of claim 135, further comprising a constructor, for constructing a graph having a plurality of paths representing the dataset.

144. The apparatus of claim 143, wherein said extractor is designed to search for partial overlaps between paths of said graph.

145. The apparatus of claim 144, wherein said graph comprises a plurality of vertices, each representing one token of the lexicon, and further wherein each path of said plurality of paths comprises a sequence of vertices respectively corresponding to one sequence of the dataset.

146. The apparatus of claim 144, further comprising electronic-calculation functionality for calculating, for each path, a set of probability functions characterizing said partial overlaps.

147. The apparatus of claim 146, wherein said extractor comprises a testing unit capable of evaluating a statistical significance of said set of probability functions.

148. The apparatus of claim 134, wherein the dataset comprises a corpus of text.

149. The apparatus of claim 134, wherein the dataset comprises a protein database.

150. The apparatus of claim 134, wherein the dataset comprises a protein database.

151. The apparatus of claim 134, wherein the dataset comprises a DNA database.

152. The apparatus of claim 134, wherein the dataset comprises an RNA database.

153. The apparatus of claim 134, wherein the dataset comprises a recorded speech.

154. The apparatus of claim 134, wherein the dataset comprises a corpus of music notes.

155. The apparatus of claim 134, wherein the dataset comprises a weblog database.

156. The apparatus of claim 134, wherein the dataset comprises trajectory records of a transportation network.

157. The apparatus of claim 134, wherein the dataset comprises activity records of a self-active system.

158. The apparatus of claim 134, wherein the dataset comprises records of operational steps in a technical process.

159. A generalized dataset produced by the method of claim 19, the generalized dataset is stored, in a retrievable and/or displayable format, on a memory medium.

160. A memory medium, storing in a retrievable and/or displayable format, the generalized dataset of claim 159.

161. A generalized dataset defined over a lexicon of tokens and stored in a retrievable and/or displayable format on a memory medium, the generalized dataset being represented by a forest hierarchy having a plurality of multilevel trees, each tree of said plurality of multilevel trees representing a pattern of tokens of the generalized dataset and comprising a leaf level, having a plurality of child nodes, and at least one partition level, having at least one parent node, wherein each child node of said leaf level corresponds to a token, and each parent node of said at least one partition level corresponds to a significant patterns of tokens or an equivalence class of tokens.

162. A memory medium, storing in a retrievable and/or displayable format, a generalized dataset defined over a lexicon of tokens and represented by a forest hierarchy having a plurality of multilevel trees, each tree of said plurality of multilevel trees representing a pattern of tokens of the generalized dataset and comprising a leaf level, having a plurality of child nodes, and at least one partition level, having at least one parent node, wherein each child node of said leaf level corresponds to a token, and each parent node of said at least one partition level corresponds to a significant patterns of tokens or an equivalence class of tokens.

163. A generalized dataset defined over a lexicon of tokens and stored in a retrievable and/or displayable format on a memory medium, the generalized dataset

being represented by a graph having a plurality of vertices selected from the group consisting of token-vertices, pattern-vertices and generalized-vertices, wherein each token-vertex represents a token of the lexicon, each pattern-vertex represents a significant pattern of tokens, and each generalized-vertex represents an equivalence class of tokens.

164. A memory medium, storing in a retrievable and/or displayable format, a generalized dataset defined over a lexicon of tokens and represented by a graph having a plurality of vertices selected from the group consisting of token-vertices, pattern-vertices and generalized-vertices, wherein each token-vertex represents a token of the lexicon, each pattern-vertex represents a significant pattern of tokens, and each generalized-vertex represents an equivalence class of tokens.